DOCUMENT RESUME

ED 167 575 TM 008 237

AUTHOR Hambleton, Ronald K.; Eignor, Daniel R.
TITLE Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals.
PUB DATE Mar 78
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Elementary Secondary Education; *Evaluation Criteria; *Specifications; *Test Construction; Testing; *Test Reviews; *Test Selection

ABSTRACT

Guidelines for evaluating criterion-referenced tests and test manuals are proposed and applied to a sample of popular commercially published tests. The well-known Test Standards published by a joint committee of professional societies is helpful, though not completely applicable, and was used together with other sources in the preparation of an evaluation form. This form is designed to be useful to both users and developers of criterion referenced tests. The 39 guideline questions were applied to 11 tests. Among the common weaknesses found were: (1) lack of domain specifications; (2) no indication of the qualifications of the individuals who prepared the test objectives; (3) possible content bias due to the use of item analysis in test construction; (4) inadequate information about test reliability; (5) lack of information about the rationale for cutting scores; (6) not enough information about error in test scores; and (7) no information about factors affecting the validity of scores. Suggestions for improving the guidelines are encouraged. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Ron Hambleton
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Guidelines for Evaluating Criterion-Referenced Tests and Test Manuals[1,2,3]

Ronald K. Hambleton and Daniel R. Eignor
University of Massachusetts, Amherst

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED167575

Most of the major test publishers have published in the last few years a wide assortment of criterion-referenced tests. In addition, many school districts, state agencies, small testing firms, and consulting firms have produced their own criterion-referenced tests. Criterion-referenced tests are designed to address many problem areas. For example, criterion-referenced tests are being used to monitor student progress through school programs, to diagnose learning disabilities, to report student progress to parents, to evaluate various types of programs, and to certify or license professionals in many fields. Unfortunately, it appears to us, and to many users of criterion-referenced tests we have spoken with, that many of the available tests fall short of the technical quality necessary for them to accomplish their intended purposes. Perhaps one explanation is that many criterion-referenced tests were developed before an adequate testing technology was fully explicated. Fortunately, there now exists an adequate technology for constructing criterion-referenced tests and using criterion-referenced test scores (Hambleton and Eignor, 1978; Hambleton, Swaminathan, Algina, Coulson, 1978; Popham, 1978). Another possible explanation is that there has been a shortage of guidelines for constructing and using criterion-referenced tests. Certainly the well-known Test Standards for

TM 008 237

[1]Paper presented at the annual meeting of NCME, Toronto, 1978.

[2]Laboratory of Psychometric and Evaluative Research Report No. 73. Amherst, MA: School of Education, University of Massachusetts, Amherst, 1978.

[3]A shorter version of this paper will appear in the Journal of Educational Measurement, 1978, in press.

evaluating tests and test manuals prepared by a joint committee of AERA/APA/NCME is helpful, but it is not completely applicable to criterion-referenced tests. Besides the incompleteness of the AERA/APA/NCME Test Standards for evaluating criterion-referenced tests and test manuals, what relevant information there is, is scattered through 75 pages or so of other materials appropriate for norm-referenced test evaluations. Therefore, the Test Standards in its present form, is not very useful for individuals interested in evaluating criterion-referenced tests.

The primary purpose of this paper is to propose a set of guidelines for evaluating criterion-referenced tests and test manuals. The guidelines should be useful to both users and developers of criterion-referenced tests. Test standards are not offered in the paper (an example of a standard is, "test score reliability must exceed .80"), but we do offer a set of questions for consideration by potential users and developers of criterion-referenced tests. The only other efforts we are aware of to develop guidelines for evaluating criterion-referenced tests and test manuals are Popham (1978, Chapter 8), Swezey and Pearlstein (1975), and Walker (1977). A secondary purpose is to report on our use of the guidelines with eleven commercially available criterion-referenced test batteries.

One caution and one comment seem appropriate to introduce at this point. The guidelines represent our own biases about what is important technical information for users to have in making informed decisions about the quality of criterion-referenced tests. Also, in this paper we did not provide (1) a rationale for the inclusion of each guideline, and (2) specifics on how the guidelines were applied. Interested readers

are encouraged to read Eignor (1978) and Hambleton and Eignor (1978) for the information.

A Proposed Set of Guidelines

The list of guidelines was generated by placing ourselves in the role of potential purchasers of a criterion-referenced test, and asking "What questions would we want to answer before making a decision to use a criterion-referenced test in a particular situation?" Questions were organized around ten broad categories[1]. They are: Objectives, Test Items, Administration, Test Layout, Reliability, Cut-off Scores, Validity, Norms, Reporting of Test Score Information, and Test Score Interpretations. The questions are as follows:

Objectives

A.1 Is the purpose (or purposes) of the test stated in a clear and concise fashion?

A.2 Is each objective clearly written so that it is possible to identify an "item pool"?

A.3 Is it clear from the list of objectives what the test measures?

A.4 Is an appropriate rationale offered for including each objective in the test?

A.5 Can a potential user "tailor" the test to meet local needs by determining which objectives from a pool of objectives offered by the publisher are to be measured by the test?

A.6 Is there a match between the content measured by the test and the situation where the test is to be used?

A.7 Are individuals identified who were responsible for the preparation of objectives?

A.8 Does the set of objectives measured by the test serve as a representative set from some content domain of interest?

---

[1]The very important factors of cost and time limits are not considered here, but they are included in our evaluation form.

B. Test Items

B.1 Is the item review process described?

B.2 Are the test items valid indicators of the objectives they were developed to measure?

B.3 Is the set of test items measuring an objective representative of the "pool" of items measuring the objective?

B.4 Are the items free of technical flaws?

B.5 Are the test items in an appropriate format to measure the objectives they were developed to measure?

B.6 Are the test items free of bias (for example, sex, ethnic, or racial)?

B.7 Was a heterogeneous sample of examinees employed in piloting the test items?

B.8 Was the item analysis data used only to detect "flawed" items?

C. Administration

C.1 Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?

C.2 Are the test directions clear?

C.3 Is the test easy to score?

C.4 Does the test manual specify an examiner's role and responsibilities?

D. Test Layout

D.1 Is the layout of the test booklets attractive?

D.2 Is the layout of the test booklets convenient for examinees?

E. Reliability

E.1 Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?

E.2 Was the sample (or samples) of examinees used in the reliability study adequate in size, and representative of the population for whom the test is intended?

E.3 Are test lengths suitable to produce tests with desirable levels of test score reliability?

E.4 Is reliability information offered in the test manual for each intended use (or uses) of the test scores?

F. Cut-Off Scores

F.1 Was a rationale offered for the selection of a method for determining cut-off scores?

F.2 Was the procedure for implementing the method explained, and was it appropriate?

F.3 Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?

G. Validity

G.1 Does the validity evidence offered in the test manual address adequately the intended use (or uses) of scores obtained from the test?

G.2 Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?

H. Norms

H.1 Are the norms data reported in an appropriate form?

H.2 Are the samples of examinees utilized in the norming study described?

H.3 Are appropriate cautions introduced for proper test score interpretations?

I. Reporting of Test Score Information

I.1 Are the test scores reported for examinees on an objective by objective basis?

I.2 Are there multiple options available to the user for reporting of test results (for example, by class and grade within a school)?

I.3 Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?

J. Test Score Interpretations

J.1 Are suitable cautions included in the manual for interpreting individual and group objective score information?

J.2 Are appropriate guidelines offered in the manual for utilizing test scores to make descriptive statements, instructional decisions, program evaluation decisions, or other stated uses of the test scores?

A convenient rating form is given on the next four pages.

Evaluation of Eleven Criterion-Referenced Tests

Eleven of the more popular criterion-referenced tests were selected for review. The names of the tests and some descriptive information are presented in the chart.

INSERT THE CHART ABOUT HERE.

Our primary purpose was to ascertain the extent to which these tests met our guidelines. We have reported our evaluation of each test relative to each guideline, but the more important information is arrived at by determining how well the tests as a group meet each of our guidelines. The group information is informative because it helps to pin-point areas where commercial materials are in need of revisions and further development.

Criterion-Referenced Test and Test Manual Evaluation Form

Background Information

Test Name:_____ Forms and Levels:_____

Test Publisher:_____ Author(s):_____

Year of Publication:_____ Cost:_____

Reusable Booklets: Yes No

Special Test Administration Conditions:_____

Manual and Other Technical Aids:_____

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question.

Question	Ratings: Acceptable	Unacceptable	Unsure	Not Applicable	Comments
A.1. Is the purpose (or purposes) of the test stated in a clear and concise fashion?					
A.2. Is each objective clearly written so that it is possible to identify an "item pool"?					
A.3. Is it clear from the list of objectives what the test measures?					
A.4. Is an appropriate rationale offered for including each objective in the test?					
A.5. Can a user "tailor" the test to meet local needs by selecting objectives from a pool of available objectives?					
A.6. Is there a match between the content measured by the test and the situation where the test is to be used?					

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "√" in the column corresponding to your answer to each question. Question	Ratings: Acceptable	Unacceptable	Unsure	Not Applicable	Comments
A.7. Are individuals identified who were responsible for the preparation of objectives?					
A.8. Does the set of objectives measured by the test serve as a representative set from some content domain of interest?					
B.1. Is the item review process described?					
B.2. Are the test items valid indicators of the objectives they were developed to measure?					
B.3. Is the set of test items measuring an objective representative of the "pool" of items measuring the objective?					
B.4. Are the items free of technical flaws?					
B.5. Are the test items in an appropriate format to measure the objectives they were developed to measure?					
B.6. Are the test items free of bias (for example, sex, ethnic, or racial)?					
B.7. Was a heterogeneous sample of examinees employed in piloting the test items?					
B.8. Was the item analysis data used only to detect "flawed" items?					
C.1. Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?					

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question. Question	Ratings: Acceptable	Unacceptable	Unsure	Not Applicable	Comments
C.2. Are the test directions clear?					
C.3. Is the test easy to score?					
C.4. Does the test manual specify an examiner's role and responsibilities?					
D.1. Is the layout of the test booklets attractive?					
D.2. Is the layout of the test booklets convenient for examinees?					
E.1. Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?					
E.2. Was the sample of examinees adequate in size, and representative of the population for whom the test is intended?					
E.3. Are test lengths suitable to produce tests with desirable levels of test score reliability?					
E.4. Is reliability information offered in the test manual for each intended use (or uses) of the test scores?					
F.1. Was a rationale offered for the selection of a method for determining cut-off scores?					
F.2. Was the procedure for implementing the method explained, and was it appropriate?					

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question. Question	Ratings: Acceptable	Unacceptable	Unsure	Not Applicable	Comments
F.3. Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?					
G.1. Does the validity evidence offered in the test manual address adequately the intended use (or uses of scores) obtained from the test?					
G.2. Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?					
H.1. Are the norms data reported in an appropriate form?					
H.2. Are the samples of examinees utilized in the norming study described?					
H.3. Are appropriate cautions introduced for proper test score interpretations?					
I.1. Are the test scores reported for examinees on an objective by objective basis?					
I.2. Are there multiple options available to the user for reporting of test results (for example, by class and grade within a school)?					
I.3. Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?					
J.1. Are suitable cautions included in the manual for interpreting individual and group objective score information?					
J.2. Are appropriate guidelines offered for utilizing test scores to accomplish stated purposes?					

Tests Selected for Review

Code	Name of Test	Grades	Levels	Forms	Publication Date	Publisher
1	1976 Stanford Diagnostic Mathematics Test	1-12	4	2	1976	Harcourt Brace Jovanovich
2	1976 Stanford Diagnostic Reading Test	1-12	4	2	1976	Harcourt Brace Jovanovich
3	Skills Monitoring System-Reading	3-5	3	1	1975	Harcourt Brace Jovanovich
4	Individual Pupil Monitoring System-Mathematics	1-6	6	2	1974	Houghton-Mifflin
5	Individual Pupil Monitoring System-Reading	1-8	8	2	1974	Houghton-Mifflin
6	Diagnostic Mathematics Inventory	1.5-7.5	7	1	1977	CTB/McGraw-Hill
7	Prescriptive Reading Inventory	K-6.5	6	1	1977	CTB/McGraw-Hill
8	Diagnosis: An Instructional Aid-Mathematics and Reading	1-6	2	2	1974	Science Research Associates
9	Mastery: An Evaluation Tool-SOBAR Reading	K-9	10	2	1975	Science Research Associates
10	Mastery: An Evaluation Tool-Mathematics	K-8	9	2	1974	Science Research Associates
11	Fountain Valley Support System in Mathematics	K-8	9	1	1974	Richard L. Zweig Associates

In judging the quality of a test and test manual relative to each guideline, the following rating scale was used:

A	= Acceptable
A'	= Acceptable, with reservations
X	= Unacceptable, data offered was unsuitable or improperly used
Y	= Unacceptable, no data was offered
N	= Not Applicable

Table 1 summarizes our ratings of the 11 tests on the 39 guidelines.

INSERT THE TABLE ABOUT HERE

Our most significant impressions of the test and test manuals reviewed are as follows:

1. In areas such as Administration, Test Layout, and Norms, there are few problems.

2. Current commercially available "criterion-referenced tests" reviewed in this paper should be called "objectives-referenced tests" since the tests appear to be developed from behavioral objectives (Popham, 1978). Starting to develop a test from a listing of behavioral objectives is less than ideal because behavioral objectives usually do not lead to unambiguous definitions of the "item pools" keyed to the behavioral objectives. The solution is to write "domain specifications" (Popham, 1978).

3. Only about half of the publishers included information about the qualifications of individuals who prepared the objectives measured by their test. The qualifications of participants in this aspect of the test development process is important information for potential users.

Table 1

Summary of Ratings of the Criterion-Referenced Tests

Question	Test 1	2	3	4	5	6	7	8	9	10	11
A1	A	A	A	A⁻	A⁻	A	A	A	A	A	X
A2	X	X	X	X	X	X	X	X	X	X	X
A3	A	A	A⁻	A	A	A	A	A	A	A	A
A4	A	A	A	A⁻	A⁻	A	A	A	A	A	X
A5	A⁻	A⁻	A	A	A	X	X	A	A	A	A
A6	A	A	A	A	A	A	A	A	A	A	A
A7	Y	Y	A⁻	Y	Y	Y	A⁻	A	A⁻	A⁻	A⁻
A8	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻	A⁻
B1	X	X	A	A⁻	A⁻	X	A⁻	Y	A	A	Y
B2	A⁻	A⁻	A	A⁻	A⁻	?[1]	A⁻	A⁻	A	A	A⁻
B3	X	X	X	X	X	X	X	X	X	X	X
B4	A	A	A	A	A	A	A	A	A	A	A
B5	A	A	A	A	A	A	A	A	A	A	A
B6	A	A	A	Y	Y	?	Y	Y	Y	A	Y
B7	A	A	A	A	A	A	A	Y	Y	Y	Y
B8	X	X	A	X	X	X	A⁻	Y	X	X	Y
C1	A	A	A	A	A	?	A	A	A	A	?[2]
C2	A	A	A	A	A	?	A	A	A	A	A
C3	A	A	A	A	A	?	A	A	A	A	A
C4	A	A	A	A	A	?	A	A	A	A	A
D1	A	A	A	A	A	?	A	A	A	A	A
D2	A	A	A	A	A	?	A	A	A	A	A
E1	A⁻	X	A⁻	Y	Y	X	X	Y	X	X	Y
E2	A	A	A	Y	Y	A	A	Y	A	A	Y
E3	A⁻	A⁻	A⁻	A⁻	A⁻	X	X	X	X	X	A⁻
E4	A⁻	A⁻	A⁻	Y	Y	X	X	Y	X	X	Y
F1	A	A	A	Y	A⁻	Y	A	X	A	A	Y
F2	A	A	X	Y	Y	X	X	Y	A	A	Y
F3	A	A	A⁻	Y	Y	Y	A⁻	Y	A⁻	A⁻	Y
G1	A	A	A	X	X	A	A	X	A⁻	A⁻	Y
G2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
H1	A	A	N	N	N	A⁻	A	N	N	N	N
H2	A	A	N	N	N	?	Y	N	N	N	N
H3	A	A	N	N	N	Y	Y	N	N	N	N
I1	A	A	A	A	A	?	A	A	A	A	A
I2	A	A	A	A	A	?	A	A	A	A	A
I3	A	A	A	A	A	?	A	A	A	A	A
J1	A⁻	A⁻	A	Y	Y	?	A⁻	Y	A⁻	A⁻	Y
J2	A	A	A	X	X	?	A	A⁻	A⁻	A⁻	A

[1]We did not have the proper materials to assess the quality of the test in the areas marked by a "?".

[2]The information was on a cassette. We did not listen to the tape and so we were not in a position to rate this aspect of the test.

4. Since test developers have not used "domain specifications", it is impossible to assess "item representativeness". Item representativeness is essential if users desire to use objective scores to "generalize to the domains of behaviors defined by the objectives." If item representativeness is not established, scores can only be interpreted in terms of the specific items included in the test.

5. "Item analysis" is an area in which there are two problems: (a) Too little explanation is offered of the choice of particular item statistics and of the specifics of item statistics usage, and (b) item statistics are used in test construction thereby "biasing" the content validity of the test in unknown ways.

6. Test score reliability was not handled very well in most of the manuals. Either (a) inappropriate information relative to the stated uses of the test scores was offered, or (b) no information was offered.

7. Cut-off scores are typically offered, but there is no rationale offered for setting cut-off scores. Procedures used for setting cut-off scores are not explained, nor is any evidence offered for the "validity" of cut-off scores (for example, do those examinees classified as "masters" typically perform better than "non-masters" on some appropriately chosen external criterion measure?).

8. Factors affecting the validity of scores are not offered in any of the manuals.

9. Only a few of the manuals introduced the notion of "error" in test scores. It is extremely important for users to have some indication of the "stability" of their objective scores and/or "consistency of mastery/non-mastery decisions".

Concluding Remarks

Our proposed guidelines were developed after careful study of the criterion-referenced testing literature and the Test Standards. However, they are offered here only to serve as a "catalyst" for further discussion and debate on a topic of considerable importance to the test and measurement field. Our use of the proposed guidelines to evaluate eleven criterion-referenced tests was intended to (1) demonstrate that the proposed

guidelines were workable, and (2) highlight areas where considerably more (or different) work on the part of test developers is needed.

Our goal for preparing this paper has been accomplished if (1) it stimulates others to extend and improve upon our guidelines, and (2) it helps to direct test developers toward more acceptable practices of criterion-referenced test construction and preparation of test manuals.

Individuals with suggestions for improving the guidelines are encouraged to write the authors.

References

Eignor, D. R. Methodological and psychometric contributions to criterion-referenced testing technology. Unpublished doctoral dissertation, University of Massachusetts, Amherst, 1978.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1978.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Swezey, R. W., & Pearlstein, R. B. Guidebook for developing criterion-referenced tests. A report prepared for the U.S. Army Research Institute for the Behavioral and Social Sciences. Reston, Virginia: Applied Science Associates, August, 1975.

Walker, C. B. Standards for evaluating criterion-referenced tests. Los Angeles: Center for the Study of Evaluation, UCLA, 1977. (Unpublished manuscript.)